

USER PROFILE CLASSIFICATION BY WEB USAGE ANALYSIS

INVENTORS

Eytan Adar
Lada A. Adamic
Francine R. Chen

CERTIFICATE OF MAILING BY "EXPRESS MAIL"
UNDER 37 C.F.R. §1.10

"Express Mail" mailing label number: EL504216709US
Date of Mailing:

I hereby certify that this correspondence is being deposited with the United States Postal Service, utilizing the "Express Mail Post Office to Addressee" service addressed to Commissioner for Patents, Washington, DC 20231 and mailed on the above Date of Mailing with the above "Express Mail" mailing label number:

Mary Beth Stone (Signature)
Mary Beth Stone
Signature Date: November 2, 2001

USER PROFILE CLASSIFICATION BY WEB USAGE ANALYSIS

INVENTORS

Eytan Adar
Lada Adamic
Francine Chen

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0001] The present invention relates to technology for determining user information through analysis of web pages accessed by a user.

2. Description of the Related Art

[0002] As Internet usage continues to rise, it becomes increasingly important to identify the demographic characteristics of Internet users. Such characteristics can help businesses and advertisers provide services to Internet users in particular demographic groups and to attract and retain new customers. To obtain this demographic information, web sites may request Internet users to enter personal demographic information. However, such user-entered information may be incomplete, thus preventing a business from obtaining a full demographic picture of a given Internet user. In other cases, demographic information supplied by an Internet user may be false or mistakenly incorrect.

[0003] Prior art machine learning techniques attempt to extrapolate user demographic information. Examples of such prior art techniques include the use of neural networks or Bayesian approaches to data extrapolation. These techniques often require excessively large amounts of computation in order to extrapolate meaningfully accurate demographic information. Such cumbersome tradeoffs thus limit the desirability of such prior art methods.

SUMMARY OF THE INVENTION

[0004] The present invention, roughly described, provides methods and systems that can be used to extrapolate user profile information from web usage. Demographic information of a test user can be predicted based on an analysis of a pattern of web pages accessed by the test user.

[0005] One embodiment of the present invention includes the step of detecting a set of web pages accessed by a test user. The accessed web pages are mapped to a first data structure. A second data structure identifies web page access patterns of users with a shared user profile attribute. A user profile attribute is assigned to the test user based on a comparison of the data structures.

[0006] In another embodiment, bias values are assigned to a set of web pages. Web pages accessed by a test user are detected. Bias values of the detected web pages are combined to obtain a combination result. A user profile attribute is assigned to the test user based on the combination result.

[0007] In a further embodiment, a set of expectation and maximization parameters are initialized. An expectation maximization process is performed using the parameters to obtain an expectation maximization process result. User profile attributes are assigned to a batch of test users in response to the expectation maximization process result.

[0008] In another embodiment, a first expectation maximization process is used to incrementally train a classifier with a set of users, each user having at least one known profile attribute. A second expectation maximization process is performed to "fold in" test user data and obtain an expectation maximization

process result. A user profile attribute is assigned to the test user in response to the expectation maximization process result.

[0009] In a further embodiment, a vector classification result, bias classification result, and probabilistic classification result are obtained. At least two of the results are combined to generate a combination result. A user profile attribute is assigned to the test user in response to the combination result.

[0010] The present invention can be implemented using hardware, software, or a combination of both hardware and software. The software used for the present invention can be stored on one or more processor readable storage devices including hard disk drives, CD-ROMs, optical disks, floppy disks, tape drives, RAM, ROM, or other suitable storage devices. In alternative embodiments, some or all of the software can be replaced by dedicated hardware including custom integrated circuits, gate arrays, FPGAs, PLDs, and special purpose computers. Hardware that can be used for the present invention includes computers, handheld devices, telephones (e.g. cellular, Internet enabled, digital, analog, hybrids, and others), and other hardware known in the art. Some of these devices include processors, memory, nonvolatile storage, input devices, and output devices.

[0011] These and other advantages of the present invention will appear more clearly from the following description in which the preferred embodiment of the invention has been set forth in conjunction with the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] Figure 1 is a block diagram depicting components of a computing system that can be used with the present invention.

[0013] Figure 2 is a block diagram depicting linked pages accessible by a user.

[0014] Figure 3 is a flow chart describing a process for determining user profile attributes through a vector comparison.

[0015] Figure 4 is a flow chart describing a process for generating a user path vector.

[0016] Figure 5 is a flow chart describing a process for generating a centroid vector.

[0017] Figure 6 is a plot illustrating numbers of user accesses per web page as measured in a sample data set.

[0018] Figure 7 is a plot illustrating accuracy rates achieved by an embodiment of the present invention using a vector comparison.

[0019] Figure 8 is a flow chart describing a process for determining user profile attributes through alternate vector comparisons.

[0020] Figure 8A illustrates a grouping of users by a convex hull drawn around training data set points having common profile attributes.

[0021] Figure 8B illustrates a grouping of users by a line separating training data set points having different profile attributes.

[0022] Figure 8C illustrates a grouping of users by a straight line approximation drawn through training data set points having different profile attributes.

[0023] Figure 9 is a flow chart describing a process for determining user profile attributes through an analysis of web page biases.

[0024] Figure 10 is a flow chart describing an expectation maximization process for determining user profile attributes.

5 [0025] Figure 11 is a flow chart describing an incremental classifier process for determining user profile attributes.

[0026] Figure 12 is a flow chart describing a batch classifier process for determining user profile attributes.

10 [0027] Figure 13 is a plot illustrating accuracy rates achieved by an embodiment of the present invention using a probabilistic latent variable analysis with a single classifier.

[0028] Figure 14 is a plot illustrating accuracy rates achieved by an embodiment of the present invention using a probabilistic latent variable analysis with a minimum threshold.

15 [0029] Figure 15 is a plot illustrating accuracy rates achieved by an embodiment of the present invention using a probabilistic latent variable analysis with stepped classifiers.

[0030] Figure 16 is a plot illustrating accuracy rates achieved by an embodiment of the present invention using a probabilistic latent variable analysis
20 with a minimum threshold and stepped classifiers.

DETAILED DESCRIPTION

[0031] When accessing a set of web pages, Internet users that share a common profile attribute, such as a particular demographic characteristic, may

choose to access similar or identical pages within the set. For example, some web pages may appeal to persons having a particular gender. However, a user having the particular gender will not necessarily access all web pages that are of interest to other users sharing the same gender. Thus, the fact that a user has accessed a particular web page can be informative, but the fact that the user has not accessed other web pages may not necessarily be as informative. In accordance with the present invention, the set of web pages accessed (or "visited") by a user comprise a web page access pattern which can be analyzed to predict profile attributes of the user.

[0032] Figure 1 illustrates a block diagram of a computer system 40 which can be used for the components of the present invention. The computer system of Figure 1 includes a processor unit 50 and main memory 52. Processor unit 50 may contain a single microprocessor, or may contain a plurality of microprocessors for configuring the computer system as a multi-processor system. Main memory 52 stores, in part, instructions and data for execution by processor unit 50. When the present invention is wholly or partially implemented in software, main memory 52 can store the executable code when in operation. Main memory 52 may include banks of dynamic random access memory (DRAM), high speed cache memory, as well as other types of memory known in the art.

[0033] The system of Figure 1 further includes a mass storage device 54, peripheral devices 56, user input devices 60, portable storage medium drives 62, a graphics subsystem 64, and an output display 66. For purposes of simplicity, the components shown in Figure 1 are depicted as being connected via a single

bus 68. However, as will be apparent to those skilled in the art, the components may be connected through one or more data transport means. For example, processor unit 50 and main memory 52 may be connected via a local microprocessor bus, and the mass storage device 54, peripheral devices 56, portable storage medium drives 62, and graphics subsystem 64 may be connected via one or more input/output (I/O) buses. Mass storage device 54, which may be implemented with a magnetic disk drive, optical disk drive, as well as other drives known in the art, is a non-volatile storage device for storing data and instructions for use by processor unit 50. In one embodiment, mass storage device 54 stores software for implementing the present invention for purposes of loading to main memory 52.

[0034] Portable storage medium drive 62 operates in conjunction with a portable non-volatile storage medium, such as a floppy disk, to input and output data and code to and from the computer system of Figure 1. In one embodiment, the system software for implementing the present invention is stored on such a portable medium, and is input to the computer system via the portable storage medium drive 62. Peripheral devices 56 may include any type of computer support device, such as an input/output (I/O) interface, to add additional functionality to the computer system. For example, peripheral devices 56 may include a network interface for connecting the computer system to a network, as well as other networking hardware such as modems, routers, and other hardware known in the art.

[0035] User input devices 60 provide a portion of a user interface. User input devices 60 may include an alpha-numeric keypad for inputting

alpha-numeric and other information, or a pointing device, such as a mouse, a
trackball, stylus, or cursor direction keys. In order to display textual and graphical
information, the computer system of Figure 1 includes graphics subsystem 64 and
output display 66. Output display 66 may include a cathode ray tube (CRT)
5 display, liquid crystal display (LCD) or other suitable display device. Graphics
subsystem 64 receives textual and graphical information, and processes the
information for output to display 66. Additionally, the system of Figure 1
includes output devices 58. Examples of suitable output devices include
speakers, printers, network interfaces, monitors, and other output devices known
10 in the art.

[0036] The components contained in the computer system of Figure 1 are
those typically found in computer systems suitable for use with certain
embodiments of the present invention, and are intended to represent a broad
category of such computer components known in the art. Thus, the computer
15 system of Figure 1 can be a personal computer, workstation, server,
minicomputer, mainframe computer, or any other computing device. Computer
system 40 can also incorporate different bus configurations, networked platforms,
multi-processor platforms, etc. Various operating systems can be used including
Unix, Linux, Windows, Macintosh OS, Palm OS, and other suitable operating
20 systems. It will also be appreciated that the present invention can be
implemented using multiples of all or parts of computer system 40 depicted in
Figure 1

[0037] Figure 2 provides a high level block diagram 100 depicting linked
web pages of one or more web sites accessible by an Internet user. In diagram

100, separate web pages are represented by nodes A, B, C, D, E, N, and Z. The nodes of diagram 100 are linked together, allowing an Internet user to trace a path from page to page using the links found at each node. In Figure 2, the page represented by node A allows the user to follow a link directly to node C or node B. However, direct links may not always be available. For example, if a user viewing the page represented by node A wishes to link to the page of node N, the user must first link to node B, and then perform a second link from node B to node N. By performing these separate links to nodes B and N, the user has traced a path from node A to node N.

[0038] In accordance with the present invention, a "user path" identifies a set of web pages accessed by a user. Thus, in the example above, the user path can be represented as: A, B, N. In an alternative notation, the user path can be represented as: A:B:N. Each web page in a user path can be identified by, among other things, IP addresses, sequentially numbered values, or positions in a web portal hierarchy of pages. In the case of a hierarchical directory service, a given web page can be identified by the user path traced from a high level page (such as the page represented by node A) to the given page.

[0039] A classification system ("classifier") in accordance with the present invention can detect web pages that have been accessed by a user. In one embodiment, this detection is performed by evaluating cookies stored by the user's web browser. Web pages that are referenced by the stored cookies are presumed to have been accessed by the user and are thus detected. In such an embodiment, cookies must be enabled on a user's web browser. In an alternate embodiment, web pages that are cached locally by a user's computer system are

ignored by the classifier. In another embodiment, web pages are deemed to be accessed by a user when viewed by the user, regardless of where the web pages are stored such as on a web server, proxy server, cached locally by a user's computer system, or elsewhere. In yet another embodiment, detection of web pages is performed by ascertaining an IP address of a user and noting which web pages are accessed from the user's IP address.

[0040] Internet users may access a particular web page multiple times. For example, if node N contains a list of useful resources (such as a list of resources provided by a directory service) available on other web pages not illustrated in Figure 2, an Internet user may choose to link back and forth between node N and the other web pages pointed at by node N. If the user's visits to these other pages are not detected or are ignored, each visit to node N can be recorded as a separate entry in a user path with no intervening user path entries. For example, if a user first accesses node A, links to node B, links to node N, links to an ignored page, and then links back to node N, the user path can be represented by: A, B, N, N. The individual web pages of a user path can also be represented as tuples. These tuples can comprise an identifier for an accessed page and the number of times that the page appears in the user path. Thus, a user path comprising the nodes: A, B, N, A, N can be represented by tuples: (A, 2), (B, 1), and (N, 2).

[0041] In accordance with the present invention, multi-dimensional vectors can be used to facilitate the determination of user profile attributes, wherein web pages are mapped to each vector dimension (or "vector index"). A user path vector is one such vector wherein the value of each vector index

corresponds to the number of times the particular web page corresponding to the vector index appears in the user path of a particular user. In one embodiment, a user path vector exists in an n-dimensional space, with each dimension corresponding to a web page, wherein visits to the web page are to be detected in accordance with the present invention. For example, referring to diagram 100 of Figure 2, if visits to the web pages at nodes A, B, C, D, E, N, and Z are to be detected, then the user path vector can be represented as: [A B C D E N Z] with a separate index for each page. In one embodiment, the value at each index of the vector is the number of times a user has accessed the web page corresponding to each particular index. Thus, applying the user path vector representation above, a user path of: A, B, N, A, N can be represented as a user path vector: [2 1 0 0 0 2 0]. A centroid vector is another multi-dimensional vector wherein the value of each vector index is determined by evaluating a set of user path vectors of Internet users having one or more known profile attributes, as further described herein. User path vectors as well as centroid vectors can be represented as data structures capable of being processed by a computer.

[0042] Figure 3 provides a flow chart 120 describing a process for determining a profile attribute of a user whose profile attribute is unknown or doubted ("test user"). In step 125, centroid vectors are generated for different values of user profile attributes, as further described herein. In step 130, a user path vector is generated for the test user, as further described herein. In step 135, the centroid vectors are compared with the user path vector. In step 137, a value for the test user's profile attribute is predicted based on the comparison of step 135. In step 140, the predicted profile attribute is assigned to the test user.

[0043] Figure 4 provides a flow chart 190 describing a process for generating a user path vector. In one embodiment, the process of Figure 4 is called by step 130 of Figure 3. In step 195, web pages accessed by a test user are detected. In step 197, a user path is identified based on the detected web pages, as previously described above. At step 200, the user path of the test user is mapped into a user path vector V .

[0044] To facilitate comparison of the user path vector mapped in step 200 with one or more centroid vectors as further described herein, optional steps 205, 210, and 215 can be performed. In some cases, certain Internet users may access many more web pages than other users. In order to minimize the effects of different numbers of web page visits between different test users while still considering the distribution of a test user's web page visits, the user path vector V mapped in step 200 can be normalized in step 205 to generate a normalized user path vector V' . In one embodiment, the normalized user path vector V' is generated as follows:

$$V'_k = \frac{V_k}{V_{\max}}$$

for each index k in the range 0 to size (V), where V_{\max} is the index having the highest value in user path vector V .

[0045] In addition to possible differences in the relative number of web pages accessed by various Internet users, certain web pages may be accessed much more frequently than other web pages when measured over many users. This difference in frequency is illustrated in plot 240 of Figure 6 which illustrates the number of user visits per web page as measured in a sample data set. As

indicated by plot 240, certain web pages in the range of page 1 to page 3,440 are accessed much more frequently than other pages. In some cases, the disparity between web page accesses is as large as several orders of magnitude. To dampen the effects of this disparity, user path vectors can be weighted to dampen these effects.

[0046] Referring again to Figure 4, the indices of the user path vector can be weighted in optional step 210. In one embodiment, this weighting is performed by maintaining a table T (not shown) which maps web pages to the total number of times each web page has been accessed. In one embodiment, an inverse document frequency ("IDF") weighting can be applied to the user path vector. By applying IDF, the weight of each web page k becomes:

$$\ln\left(\frac{N}{T_k}\right)$$

where N is the total number of unique users who have accessed web page k , and T_k is the total number of times web page k has been accessed.

[0047] The normalization obtained in step 205 and the page weighting obtained in step 210 can be combined to generate a normalized-weighted user path vector P in step 215. In one embodiment, the indices of P are calculated as follows:

$$P_i = V_i' \ln\left(\frac{N}{T_i}\right)$$

for each i in the range 0 to size (V). The use of P during comparison step 135 can minimize the effects of wide disparities between relative numbers of web pages accessed by different users, as well as the effects of differences in the

number of times various web pages have been accessed when measured over many users, as discussed above.

[0048] Figure 5 provides a flow chart 160 describing a process for generating a centroid vector. In one embodiment, the process of Figure 5 is called by step 125 of Figure 3. In order to generate a centroid vector, a set of user path vectors are generated for users in a sample data set for whom at least one profile attribute is known in step 163. In one embodiment, step 163 calls the process of Figure 4. If the gender of a test user is sought to be classified, then the set of user path vectors are generated from user paths of Internet users for whom a gender profile attribute is known. In step 165, the user paths of users in the sample set are separated into clusters distinguished by the value of the known attribute. Thus, if gender of a test user is to be classified, then all user paths of sample set users known to be male can be placed in one cluster, and the remaining user paths of sample set users known to be female can be placed in a second cluster. This cluster grouping facilitates the generation of separate centroid vectors for male and female users in the sample set as further described herein.

[0049] In step 170, the index values of one or more centroid vectors are calculated. For example, if gender is to be classified, separate centroid vectors can be generated for the male and female clusters of sample set users. The user path of each user in the sample set can be represented as a user path vector having indices corresponding to different web pages. The number of times that a sample set user accesses a page can be represented numerically by an index of the user path vector. In one embodiment, the indices of the centroid vector for each

cluster correspond to the average values of the indices of the user path vectors generated from user paths in the corresponding cluster. For example, each index C_i of a male cluster centroid vector C^m can be calculated as follows:

$$C_i = \frac{\sum_k V_i^k}{M}$$

where V_i^k is the value at index i for the vector representing the k th male sample set user and M is the number of male users in the sample set. The indices of female cluster centroid vector C^f can be similarly calculated by substituting female values into the equation above, where V_i^k is the value at index i for the vector representing the k th female sample set user and M is the number of female users in the sample set. As a result of calculating C_i for each index of each cluster, separate multi-dimensional centroid vectors C^m and C^f are constructed.

[0050] Referring to Figure 3, after generation steps 125 and 130, vector P can be compared to centroid vectors C^m and C^f in step 135. Various distance metrics can be used to evaluate the distance between P and C^m as well as the distance between P and C^f . In one embodiment, the centroid vector having the shortest distance from vector P is predicted to correspond to a profile attribute of the test user represented by P (step 137) and assigned to the test user (step 140). In one embodiment, the distance between vector P and a centroid vector C is determined using the cosine distance:

$$\cos \theta = \frac{\sum_i P_i * C_i}{\|P\| * \|C\|}$$

Using this method, the test user is predicted to have the user attribute of the cluster for which the cosine value is the greatest. For example, if gender is the user profile attribute to be predicted, then a greater cosine value measured between P and C^m than between P and C^f would indicate that the test user's behavior more closely matches the behavior of an "average" male user than the behavior of an "average" female user. As a result, a male user will be predicted (step 137) and assigned to the test user (step 140).

[0051] Figure 7 provides a plot 260 depicting classification accuracy rates achieved by an embodiment of the present invention using a vector comparison. Plot 260 illustrates the percentage of times that a test user's gender was guessed correctly for different numbers of data samples. To generate plot 260, log files from a major Internet portal web site were used to generate centroid vectors for the gender of a sample set of users for whom gender was known. The y-axis of plot 260 measures the accuracy of predicting a correct gender user profile attribute (i.e. the number of correctly classified users divided by the total number of users guessed). The x-axis measures the number of web page accesses by the test user that were considered. As indicated by plot 260, accuracy increases as more web page visits are considered. This experimental data indicates that a classifier in accordance with the present invention can predict the gender of a test user with an accuracy of over 75% when a sufficient number of web sites are visited by the user.

[0052] In some cases, users having certain profile attributes may access a great many more web pages than persons having other attributes. For example, in the log files described above, users identifying themselves as females accessed

web pages distributed across a greater number of web sites than users identifying themselves as males. As a result, the number of non-zero indices in the female centroid vector C^f for the above data was consistently much greater than the male centroid vector C^m . Thus, the cosine value calculated in the distance metric above was always higher when using female centroid vector C^f . This caused the number of predicted females to be biased upwards. To counteract this effect, the distance between vector P and a female centroid vector C^f can be artificially increased.

This technique was applied in the experiment that generated plot 260. In one embodiment, each index of vector P can be reduced to implement this change in distance. In another embodiment, the cosine distance measured between vector P and female centroid vector C^f can be multiplied by a reducing factor (for example, 0.8).

[0053] Other refinements can further improve the accuracy of a vector comparison classifier in accordance with the present invention. The centroid vectors C^m and C^f can be truncated in a number of different ways. For example, a principal component analysis, as it is understood by those skilled in the art, can be applied to reduce the dimensions of the centroid vectors. This technique ignores indices of the vectors that are not informative, such as indices corresponding to web pages that do not bear a strong relationship to gender. For example, such a technique may cause the entry page of a web portal site to be ignored.

[0054] In the experiments described above, the test sample set used to generate the centroid vectors was artificially selected to represent an equal male/female distribution. However, real world experience may not necessarily mirror such an equal distribution. For example, if males comprise 60% of all

Internet users and females comprise 40% of all Internet users, the principles of Bayes Law, as it is understood by those skilled in the art, can be applied to take into account the *a priori* distribution.

5 [0055] Web page access patterns can depend on multiple user attributes, such as the gender as well as the occupation of a given test user. Certain occupational distributions that vary by male/female user attributes can be combined with the gender determinations above to further improve the accuracy of a classifier in accordance with the present invention.

10 [0056] The dependence of web pages to other web pages can also be considered by a classifier in accordance with the present invention. For example, in diagram 100 of Figure 2, a user's act of linking to node Z followed by a link to node N is not necessarily informative for purposes of determining profile attributes of the user. If node Z has only a single link to node N with no links to other pages, then the strong relationship between node Z and node N can create an artificially high number of accesses to node N. In such a case, the vector indices corresponding to node N can be reduced in value, or simply not considered, in order to offset the artificially high value. In another embodiment, such web page dependencies are ignored by the classifier.

15 [0057] As a further refinement, different transition probabilities for different user profile attributes can be considered. For example, if it is known that male users tend to make a particular transition from one web page to another web page while females tend to perform a different transition, this information can be instructive in the prediction of a test user's gender.

[0058] In another embodiment, alternate distance metrics can be used for calculating the distance between vector P and centroid vectors C^m and C^j .

Examples of such alternate distance metrics include counting the number of steps between the vectors using a city street distance calculation or performing a Euclidian distance calculation, as these calculations are known in the art.

[0059] The present invention can be further refined by using alternate ways of comparing vector P with centroid vectors C^m and C^j in step 135 other than, or in addition to, the distance metrics discussed above. Figures 8A - C illustrate several such alternatives. In these figures, a reduced web page space of two pages is assumed wherein a given user will access pages 1 and 2 a total of m and n times, respectively. The user path vector of the user can therefore be represented as $[M\ N]$. The vectors of users having a known gender are plotted in the two-dimensional space and marked with a point to indicate their profile attribute as male (X) or female (O). Ideally, members of the male classification would fall into a first localized area as represented in the two-dimensional space, with the female classification in a second localized area.

[0060] Figure 8 is a flow chart 262 describing a process for determining user profile attributes through alternate vector comparisons. In step 264, user path vectors are generated for sample set users. Clusters of sample set users having profile attributes in common are then identified in step 266. In step 268, a user path vector is generated for the test user. A distance is calculated between the user path vector of the test user and each identified cluster (step 270). In step 272, a user profile attribute is predicted for the test user. In one embodiment, the profile attribute associated with the cluster having the shortest distance from the

test user path vector is predicted for the test user. In step 274, the profile attribute predicted in step 272 is assigned to the test user.

5 [0061] Figure 8A illustrates the use of convex hulls 285 and 290 drawn around the clusters of users with known genders. Test users u1 and u2 can be plotted in the two-dimensional space as indicated. To test whether users u1 and u2 should be classified as male or female, a distance from each of users u1 and u2 to convex hulls 285 and 290 is measured in step 135. Each of users u1 and u2 is then predicted to have the profile attribute corresponding to the closest measured cluster/hull combination.

10 [0062] Figure 8B illustrates a grouping of users by a line 315 separating male clusters from female clusters. The gender of unknown users u1 and u2 can be determined by evaluating whether they reside on the male cluster side or the female cluster side of line 315.

15 [0063] Figure 8C provides a plot 330 illustrating a grouping of users by a straight line approximation 335 drawn through the points representing users whose gender is known. Similar to Figure 8B, the gender of unknown users u1 and u2 can be predicted by determining which side of line 335 unknown users u1 and u2 fall. Of the three techniques illustrated in Figures 8A - C, straight line approximation 335 is preferred. It can minimize the difficulties of drawing convex hulls 285 or 290 around data sets that overlap, as well as minimize the difficulties of drawing a line 315 that completely separates known male users from female users. Straight line approximation 335 further minimizes the difficulties encountered when calculating the distance between an unknown point and a dividing line.

20

[0064] Figure 9 provides a flow chart 360 describing a process for determining user profile attributes through a comparison of web page biases calculated from a sample data set. In accordance with the present invention, a user profile attribute can be determined by evaluating bias values assigned to web pages accessed by a user. The biases of all accessed pages can be summed to yield a net bias of the user. The process of Figure 9 can be used as an alternative, or in conjunction with the process of Figure 3.

[0065] In step 370, the bias of each web page visited by a test user is calculated. In one embodiment, the bias of a particular web page is the difference between: the actual number of users having a certain attribute who visit the page, and the product of the total number of users who visit the page and the fraction of users having the attribute as measured over a set of web pages that includes the particular web page. The bias can be further normalized by the expected deviation in the number of visitors from the expected value which depends on the number of visitors to the page. The gender bias b of a particular web page can be calculated as follows:

$$b = \frac{(M - m * N)}{\sqrt{N * m * (1 - m)}}$$

where m is the fraction of all users that are male as measured over a set of web pages that includes the particular web page, M is the number of males who visit the particular web page, and N is the total number of users who have accessed the particular web page. Thus, if the overall fraction of male users as measured over all web pages of a web site users is 50% ($m = 0.5$), and a given web page was

accessed by 10 users, 8 of which were male, the bias of the given web page would be positive, indicating a male bias:

$$b = \frac{(8 - 10 * 0.5)}{\sqrt{10 * 0.5 * (1 - 0.5)}} = 1.90$$

5 On the other hand, if the web page was accessed by 10 users, 4 of which were male, then the bias of the web page would be negative, indicating a female bias:

$$b = \frac{(4 - 10 * 0.5)}{\sqrt{10 * 0.5 * (1 - 0.5)}} = -0.63$$

In the examples above, the highest male or female bias for a web page which was accessed by 10 users is ± 3.16 , which would occur if all users accessing the web page were either male or female.

10 **[0066]** Applying the bias calculation to other examples, if $m = 0.5$, a web page that is accessed by 3 male and 1 female user would have a calculated bias equal to 1.0. However, if the same site is accessed by 30 male users and 10 female users, the bias would equal 3.2. Thus, it is clear that with increased numbers of users, the calculated bias of a page can increase if relative user ratios are maintained.

15 **[0067]** Referring to Figure 9, in step 375, the biases of all web pages visited by a test user are summed, yielding a net bias for the particular profile attribute sought to be determined. The unknown user profile attribute of the test user can be predicted (step 377) in accordance with the net bias determined in step 375 and assigned to the test user (step 380). Thus, using the bias assignments above, a male gender would be predicted in step 377 for the test user if the result of step 375 is positive. On the other hand, if the net bias is negative,

then a female gender would be predicted. In experiments performed using an embodiment of the bias classifier process of Figure 9, male users were predicted with a 58% accuracy rate while female users were predicted with a 61% accuracy rate.

5 **[0068]** The present invention further provides a classifier capable of performing a probabilistic latent variable analysis of web page access patterns to predict user profile attributes. A number of latent variables can be specified to correspond to a number of classes of a given user profile attribute (i.e. different gender or age bracket classes) sought to be predicted by the classifier.

10 **[0069]** In accordance with a probabilistic classifier of the present invention, the conditional probability of a particular user profile attribute given a particular test user: $P(g|u)$, can be determined. Training data to be considered by a probabilistic classifier in accordance with the present invention can be represented as sets of labeled triplets: (g,s,u) , where g is a user profile attribute sought to be determined by the classifier, s is a web page visited by a user, and u is a user selected from a uniform distribution. Similarly, test data can be represented as sets of labeled pairs: (s,u) . Given a user u , a user profile attribute can be predicted based on the conditional probability of the gender given the user: $P(g|u)$. Given a gender g , a particular web page s is accessed with probability: $P(s|g)$.

20 **[0070]** Assuming that a user's gender determines whether the user accesses a web page, the probability of a particular web page being accessed by a user u with a particular gender g : $P(s|gu)$, can be approximated as: $P(s|g)$. Thus, the probability of observing a particular labeled pair (s,u) can be approximated as:

$$P(s,u) = P(u) \sum_g P(s|g)P(g|u)$$

where $P(u)$ is the probability of choosing a particular user from a uniform distribution of users.

5 **[0071]** In accordance with a probabilistic classifier of the present invention, an expectation maximization ("EM") process performed by an instructable machine can be used to iteratively fit parameters calculated by the classifier by maximizing a log-likelihood result. See Dempster, et al., "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Statist. Soc. B 39, 1977, incorporated by reference herein.

10 **[0072]** Figure 10 provides a flow chart 440 describing an EM process. In one embodiment, the process of Figure 10 is called by steps 410 and 420 of Figure 11. In another embodiment, the process of Figure 10 is called by step 500 of Figure 12. In step 445, an expectation step is performed. In one embodiment, expectation step 445 determines $P(g|s,u)$ as follows:

$$P(g|s,u) = \frac{P(s|g)P(g|u)}{\sum_{g'} P(s|g')P(g'|u)}$$

The parameters $P(s|g)$ and $P(g|u)$ used in a first iteration of step 445 can be initialized by an initialization step performed prior to the execution of Figure 10.

20 **[0073]** In step 450, a maximization step is performed. In one embodiment, maximization step 450 determines values for $P(s|g)$ and $P(g|u)$ as follows:

$$P(s|g) = \frac{\sum_u n(s,u)P(g|s,u)}{\sum_{s'} \sum_u n(s',u)P(g|s',u)} \text{ and } P(g|u) = \frac{\sum_s n(s,u)P(g|s,u)}{\sum_{s'} \sum_{g'} n(s',u)P(g'|s',u)}$$

In one embodiment, the parameter $P(g|s,u)$ used in maximization step 450 is provided by the result of estimation step 445. The parameter $n(s,u)$ of maximization step 450 indicates the number of times user u has accessed web site s . In step 455, a log-likelihood is calculated. In one embodiment, the log-likelihood is determined as follows:

$$L = \sum_g \sum_u n(s,u) \log P(s,u)$$

In another embodiment, in step 455, the accuracy on a separate validation set of data is calculated using "folding in" to determine an accuracy value.

[0074] In step 460, the process of Figure 10 determines whether to repeat steps 445, 450, and 455. If the steps are repeated, then the values of $P(s|g)$ and $P(g|u)$ calculated during the most recent maximization step 450 are substituted as the values of $P(s|g)$ and $P(g|u)$ in the next expectation step 445. Similarly, the value of $P(g|s,u)$ calculated during the next expectation step 445 will be used in the next maximization step 450. As a result of these substitutions, the values of parameters calculated by the EM process of Figure 10 can become increasingly accurate as multiple iterations of steps 445 and 450 are performed. In one embodiment, steps 445, 450, and 455 are repeated if the log-likelihood determined in step 455 has not decreased more than a threshold amount since a previous iteration of step 455. In another embodiment, steps 445, 450, and 455 are repeated if the accuracy value determined in step 455 has not decreased more

than a threshold amount since a previous iteration of step 455. In another embodiment, the steps will be repeated until a fixed number of iterations has been performed, such as 100 iterations. If the steps are not repeated (step 465), then the process proceeds to step 465 where it returns.

5 [0075] Figure 11 provides a flowchart 400 describing an incremental classifier process employing “folding in” for determining user profile attributes. See Hofman, Thomas, “Probabilistic Latent Semantic Indexing,” Proc. SIGIR '99, pp. 50-57, 1999, incorporated by reference herein. An EM process is run using data from a training set of users having a known user profile attribute. The training set data is used to initialize parameters utilized by the EM process. As a result of the EM process, a value for the conditional probability of a web page s given a user profile attribute g is determined: $P(s|g)$. A second EM process is run to “fold in” data for a test user in order to determine a conditional probability of the classes of the user profile attribute sought to be determined, given the test user: $P(g|u)$.

10
15 [0076] In step 405, parameters for expectation and maximization steps are initialized for all sets of (g,s,u) in a training set of users for whom a user profile attribute g and accessed web pages s are known. In one embodiment, $P(s|g)$ is initialized to a value equal to: $1/(\text{number of web pages considered by the classifier})$. In another embodiment, $P(g|u)$ is initialized to a value of ϵ or $1 - \epsilon$, where ϵ is close to 0. In one embodiment, ϵ is set equal to 0.00001. In step 410, separate EM processes are performed for each set of (g,s,u) in the training set. As a result of step 410, the classifier is trained and $P(s|g)$ is determined for all sets of s and g in the training set. When the process of Figure

10 is called by step 410 of Figure 11, both parameters $P(s|g)$ and $P(g|u)$ are calculated. In step 413, web pages s accessed by a test user are detected. In step 415, new EM parameters are added to the model and initialized for all sets of (g,s,u) where u in this case is a test user whose user attribute is sought to be determined. These initializations can be performed using the values of $P(s|g)$ calculated in step 410. In one embodiment, the parameter $P(g|u)$ is initialized to a value of 0.5. In step 420, separate EM processes are performed for each set of (g,s,u) (where u is the test user in this case) using the newly initialized parameters from step 415, thus "folding in" the test user data. When the process of Figure 10 is called by step 420 of Figure 11, only parameter $P(g|u)$ for only the test user u is updated in the maximization step 450, and only $P(g|s, u)$ for u equal to the test user is updated in the expectation step 445. As a result of performing step 420, a value for $P(g|u)$ will be determined for the test user.

[0077] In accordance with the present invention, a batch classifier approach can be used to determine user profile attributes for a set of test users that are combined with a training set of users for whom user profile attributes are known. Figure 12 provides a flow chart 490 describing a batch classifier process. In step 493, web pages s accessed by one or more test users are detected. Similar to step 405 of Figure 10, step 495 of Figure 12 initializes EM parameters for separate EM processes to be run for all sets of (g,s,u) . For all users in the training set for whom gender is known, EM parameters are initialized as described above with respect to step 405. For test users for whom the sought user profile attribute is not known, these parameters are initialized as described above with respect to step 415. In step 500, separate EM processes are run on all sets of (g,s,u) . As a

result of step 500, a value for $P(g|u)$ is determined for all test users for whom the sought user profile attribute was not known.

[0078] In both the incremental and batch probability classifier processes above, a value for $P(g|u)$ is determined for each user. In one embodiment, the user profile attribute for which this parameter is greatest is predicted to be the user profile attribute of the user.

[0079] To evaluate the incremental and batch probability classifier processes above, users of a major Internet portal web site were analyzed. Table 1 below illustrates the classification results achieved by an incremental classifier process in accordance with the present invention. The incremental classifier was trained on a set of 615115 users with balanced male/female proportions, and then data for an independent balanced set of 153495 users was folded in to be classified.

Table 1				
	% Correct	% Incorrect	% Unknown	Total
Male	38	62	0	76748
Female	83	17	0	76747
Total	60	40	0	153495

[0080] Table 2 below illustrates the classification results achieved by a batch classifier process in accordance with the present invention. The batch classifier was initialized based on the labels for a balanced set of 615115 users and then initialized uniformly for the separate balanced set of 153495 users considered by the incremental classifier process above. From Tables 1 and 2, it is

apparent that the incremental and batch classifiers can achieve similar performance when using the same data set.

Table 2				
	% Correct	% Incorrect	% Unknown	Total
Male	36	64	0	76748
Female	84	16	0	76747
Total	60	40	0	153495

[0081] In a second experiment using the incremental classifier, the classifier was trained on approximately 900,000 users for whom gender was known. Males comprised 66% of the training set data. The classifier performance was evaluated for all users which had visited at least N pages (a “step”), where N ranged from 1 to 200. For example, for N equal to 1, the first page visited by each user was input to the classifier.

[0082] Figure 13 provides a plot 520 illustrating accuracy rates as a function of the number of pages visited. The male performance is labeled “m,” the female performance is labeled “f,” and the overall performance is labeled “*.” As indicated by plot 520, males are classified with a higher accuracy than females as the number of accessed pages increases. When only a small number of pages have been visited by a user, then unless the user visits one of the traditional male pages, the chances are greater that a user will visit a random page that is predominately female. This bias of a “random” page being predominantly visited by females is observed in plot 520 in that where few pages have been visited, the female accuracy rate is higher.

[0083] In a third experiment using the incremental classifier, a threshold was set. In this experiment, $P(g|u)$ must be equal or greater than the threshold in order for the classifier to predict the gender of a test user. Although the threshold can be made dependent on the user attribute class (such as a threshold of 0.99 for female probabilities and a threshold of 0.5 for male probabilities, or vice versa), a single threshold of 0.99 was used for both gender classes in this experiment.

Figure 14 provides a plot 540 illustrating accuracy rates for achieved using this single threshold. In addition to the labels used in Figure 13, the overall percentage of users for which a classification decision is made is labeled "g" in Figure 14. As indicated by plot 540, except for the case of one page access, as more pages are accessed, the number of users for which a classification decision is made (the score is above threshold) increases. At a threshold of 0.99, when one page has been visited, 45% of all users are predicted with an overall accuracy of 61%, with an accuracy of 56% and 73% for males and females, respectively. When 200 pages have been visited, then 60% of all users are predicted with an overall accuracy of 82%, and an accuracy of 88% for males and 53% for females, respectively.

[0084] In a fourth experiment, a separate incremented classifier was trained for each set of page visits. For example, for a set of N pages in the range 1 to 200, a classifier was created using the first N pages visited by each user in the training set that had visited at least N pages. Figure 15 provides a plot 560 illustrating accuracy rates achieved by multiple stepped classifiers when analyzing test data. It will be appreciated that the average performance of the multiple stepped classifier approach illustrated in Figure 15 is better than the

performance of the single classifier approach illustrated in Figure 13 when the number of web page visits is small. Figure 16 provides a plot 580 illustrating accuracy rates achieved when a threshold of 0.99 was used by the multiple stepped classifiers. In Figure 16, when 7 pages have been visited, an accuracy of approximately 80% is obtained when 27% of the users are classified. As indicated in Figure 16, the accuracy remains approximately the same and a greater percentage of users are classified as the number of accessed pages increases. Thus, of the experiments above, the use of multiple classifiers utilizing a threshold achieved the highest accuracy rates given the experimental data.

10 **[0085]** In another embodiment of the present invention, stepped classifiers are utilized in the analysis of users who have visited only a few pages, while a combined classifier is used when a larger number of pages are visited (i.e. 20 pages). Subsampling of the page visits, such as creating classifiers only for the cases when 1, 3, 5, 7, 13, and 15 pages have been visited can be used to further reduce the number of classifiers needed with this method. A user that visits 6 pages, for example, can be classified using only the first 5 pages visited. The amount of memory required by a probabilistic classifier in accordance with the present invention can be further reduced by selecting a subset of pages to use.

15 **[0086]** To improve the accuracy of the probabilistic classifier discussed above, tempering can be used to prevent overfitting of data. In one embodiment, expectation step 445 is calculated as follows:

$$P(g|s,u) = \frac{[P(s|g)P(g|u)]^B}{\sum_{g'} [P(s|g')P(g'|u)]^B}$$

where B is initialized to a value of 1 and can be reduced as desired to improve accuracy. See Hofman, Thomas, "Probabilistic Latent Semantic Indexing," Proc. SIGIR '99, pp. 50-57, 1999.

5 [0087] In another refinement, multi-class profiling can be performed for user profile attributes having several classes. Examples of such multi-class attributes include: age brackets, professions, and household income levels. The number of latent variables g can be set to the number of classes. In one embodiment, the parameter $P(g|u)$ is initialized to a value of $1 - \epsilon$ in initialization steps 405, 415, and 495, where ϵ is a number much less than 1.0. 10 In another embodiment, a threshold can be set on the parameter $P(g|u)$ such that a user profile determination is not performed unless the value of $P(g|u)$ is greater than the threshold.

[0088] In another embodiment, the number of subsets considered by the probabilistic classifier can be reduced. This can reduce the amount of memory 15 required by the classifier. For example, the average mutual information $MI(g,u)$ between a gender user profile attribute and users for each web page considered by the classifier can be determined as follows:

$$MI(g,u) = \sum_g \sum_u P(g,s,u) \log \frac{P(g|s,u)}{P(g)}$$

20 For each gender, the N users with the largest MI values are selected, where N is an integer greater than 1.

[0089] To enhance the accuracy of the vector, web page bias, and probabilistic classifiers described above, the results of all or subsets of the

classifiers can be combined in a variety of ways. For example, the results of the classifiers can be combined in a linear combination. The results can also be combined in a weighted linear fashion by multiplying each result by a factor and summing the products. Similarly, the results of each classifier can be multiplied together with coefficients, as desired. In addition, results from different classifiers can be obtained depending on the total number of web pages visited by a test user. For example, if the total number of pages falls within a first range of numbers, a first classifier can be used to predict a user profile attribute. If the total number of pages falls within a second range, a different classifier can be used as an alternative, or in addition to the first classifier.

[0090] The foregoing detailed description of the invention has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. For example, although present invention is described herein in relation to user access of Internet web pages, it will be understood that the present invention is similarly applicable to computing environments other than the Internet, as well as to the accessing of data other than web pages. The described embodiments were chosen in order to best explain the principles of the invention and its practical application to thereby enable others skilled in the art to best utilize the invention in various embodiments and with various modifications as are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the claims appended hereto.